Prediction of Recurrence and Metastasis in Breast Cancer

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies

(Deemed to be University)

by

Scholar Name: CHARANPREET KAUR

Registration number: 20/Ph.D/054

Under the supervision of

Dr. Rosy Madaan
Associate Professor, Faculty of Engineering and Technology

School of Engineering and Technology

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

Certificate

I hereby declare that the short synopsis "**Prediction of Recurrence and Metastasis in Breast Cancer**" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science Engineering under Faculty of Engineering of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2022.

It is a bona fide record of original work carried out under the supervision of **Dr. Rosy Madaan** and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

CHARANPREET KAUR

Registration number: 20/Ph.D/054

Signature of the Supervisor

DR. ROSY MADAAN
Associate Professor, CSE, FET, MRIIRS

(Chairperson, DRC)

Director, CDP

Abstract

The healthcare industry produces a massive amount of information that is intricately detailed about people and their medical conditions. Medical data can contain hidden patterns or relationships that can be uncovered using data mining techniques. The spread of the disease and recurrence causes most breast cancer fatalities. One of the most significant issues in the field of medical research is the early identification of breast cancer. It is critical to determine if a tumor is benign or malignant for deciding the kind of treatment and, more importantly, to keep medical expenditures as low as possible. Data mining can assist healthcare organizations in predicting trends in patient conditions and behaviors. This is accomplished through data analysis from multiple perspectives and the discovery of connections and relationships between seemingly unrelated pieces of information, which is achieved through data mining. A metastatic recurrence might occur months after the initial diagnosis and treatment for breast cancer. Diverse strategies to forecast breast cancer have been discovered via various research studies. A novel solution to the issue was developed using Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) pictures taken before and after treatment. This study intends to ensure that patients with malignancy receive the appropriate examination. It can better discriminate between cancerous, benign, and healthy breast parenchymal tissue, resulting in a more accurate diagnostic tool using enhanced algorithms and techniques. According to recent research, the number of risk factors for breast cancer in women is increasing at an alarming pace. Regardless of one's opinion on risk evaluation, categorizing patients who have cancer into malignant and benign subtypes has been a crucial diagnostic technique for the illness. Due to the difficulty of predicting breast cancer recurrences early during treatment, only a few strategies have been examined in state of the art. The research will focus on identifying common risk factors leading to breast cancer. It will aim to build a hybrid model to predict the recurrence of breast cancer and metastasis. Additionally, the accuracy of the model and its performance will be compared to that of conventional classifiers, and then the findings and results will be interpreted.

Keywords: ***Breast Cancer, Classification, Data Mining, Healthcare Sector, Metastasis, Prediction, Risk Factors***

CONTENTS

S.NO.	DESCRIPTION	PAGE NO.
1.	Introduction	1
2.	Literature Review	2
3.	Research Gaps	13
4.	Description of Broad Area	14
4.1	Data Mining Process	13
4.2	Data mining in healthcare	15
4.3	Diagnosis of Cancer	15
4.4	Data mining techniques for breast cancer	17
5.	Problem Statement	19
6.	Research Objectives	20
7.	Methodology to be adopted	20
8.	Expected Outcome of Research	22
9.	References	23

1. Introduction

Today, there are more than 4.66 billion active internet users worldwide, as recorded in January 2021. This constitutes a total of 59.5 percent of the global population around the world. Around 83% of internet users have increased in just over five years. Every second, Google processes more than 40,000 searches, or 3.5 billion searches a day. This means around 2 trillion searches per year worldwide. It is very clear from these data statistics that data storage is increasing exponentially. There are bulks and bulks of data generated every day, and managing this ton of data today is not an issue anymore. Specialized software tools and applications can be used to analyze the data. It is a very complex process to uncover hidden information and identify associated patterns, correlations, and market trends. The use of predictive analysis, data mining, text mining, forecasting, and data optimization is widespread for data analysis today. Data Analytics leads to the development of new business opportunities, the study of market trends, and explore the information that has not been discovered yet. “Data mining (DM) is a hidden predictive data extraction of relationships and unknown data, trends, and observations by analyzing large data sets using traditional, hard-to-find and detect statistical methods” (Poonia, 2003). Data mining is an interdisciplinary field and is considered a misnomer. Data mining is the technology that analyzes past data with the present to predict the future. The specific act of data mining, according to the American Association for Artificial Intelligence's 1996 definition of knowledge discovery in data, occurs somewhere in the middle of the data processing cycle, following selection, cleaning, and normalization but before interpretation, evaluation, and any necessary subsequent refinement of the initial query or model.

In the healthcare sector, patient data has been increasing exponentially today, which has led to the rise of Electronic Health Records (EHR). With data mining, healthcare organizations have optimized treatment efficiency, increased patient life, diagnose the disease at an early stage, customer relationship management, etc. Data mining enables medical professionals to treat patients more effectively and comprehend medical assistance better. There are many challenges faced while dealing with a large amount of data. Data mining and predictive analysis have opened new access for the diagnosis and treatment of diseases at an early stage in the Healthcare industry. Data mining in the medical field helps in evaluating and predicting early symptoms to treat any

disease or condition. This reduces the money spend and saves the lives of the patients because, usually in the medical field, the decisions are made on the basis of the doctor's experience rather than on the hidden information in the databases. Utilizing data mining algorithms effectively in the healthcare sector can improve healthcare facilities and further minimize hospital admissions. When there are strange or irregular patterns or fraudulent insurance or medical policies, data mining principles can be utilized to spot fraud and abuse.

2. Literature Review

The following investigation for Breast cancer prediction and its recurrence using Data mining techniques was studied. The proposed interpretations of different researchers and their findings are presented in the table (Table 1).

Ahmad, S., Ullah, T., et al., (2022) presented a concept of a hybrid lymph node AlexNet-gated recurrent unit (AlexNet-GRU) (LN). The study used and assessed three models, including the convolutional neural network, CNN long short-term memory (CNN-LSTM), and the suggested AlexNet-GRU, and found that the proposed model performed 3% better on precision and other efficiency measures. In future, enhancement can be done by adding more layers to the proposed model to further improve the accuracy.

Merouane, E. and Said, A., (2022) used Cross Industry Standard Process for Data Mining (CRISP-DM) deep learning to forecast metastatic relapse in breast cancer patients, enabling the delivery of the necessary care and reducing cancerous cell return. A total of 449 records from Morocco were used for the study. The performance parameters of Logistic Regression, Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN) were compared. SVM was the best model compared to others, with 90% accuracy.

Alwohaibi, M., Alzaqebah, M., et al., (2021) examined two breast cancer datasets using the Brain Storming Optimization algorithm (BSO). In the first level, the statistical feature selection techniques (SFM), that selected features based on relevance rankings and correlations, were used. There were three main steps. The second-level stage, which evaluates each strategy using different classifiers, comes next. The correct feature combination is then selected in the third stage. Applied methods included linear discrimination analysis, SVM, and logistic regression.

Magboo, V.P.C. and Magboo, M.S.A., (2021) utilized four classifiers, Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines for early cancer relapse in patients post-treatment. The treatment for cancer relapse can be proposed early based on the patient's features. A modest, high-dimensional data set that categorizes breast cancer recurrence on four distinct configurations was employed.

Okagbue, H.I., Adamu, P.I., et al., (2021) used several independent factors to predict whether patients with breast cancer would survive, including age, sex, duration of illness, Region of the tumor, and the diagnosis mode. 25 male and 175 female records were included in the dataset. There are ten different models used in this study, including stochastic gradient descent (SGD), k-nearest neighbour (kNN), support vector machine (SVM), naive Bayes (NB), adaptive boosting (AdaBoost), gradient boosting (GraBoost), random forest (RF), classification tree (Tree), and logistic regression (LR). Out of all these, AdaBoost performed the best in terms of accuracy.

Yang, P.T., Wu, W.S., et al., (2021) assessed that a breast cancer relapse might be fatal and more severe. To forecast the disease's recurrence, the scientists presented a model based on clinical and numerical data. The Breast Cancer Registry, having 1061 entries, served as the source of information. The methodology consisted of three stages; preprocessing and feature selection were done in the first stage. Resampling was done in the next step, and traditional classifier techniques were finally employed in the third stage to predict breast cancer recurrence. Utilizing AdaBoost's three-fold cross-validation and cost-sensitive learning methods, performance was evaluated.

Kaur, I., Doja, M.N., et al., (2021) implemented the research to study the mortality rate for ovarian cancer patients. Sequence mining techniques were used to follow the treatment given to the patients and to create time periods for the treatment data. The Ensemble technique was used to classify the patients as survived or deceased. A Dataset of 140 ovarian cancer patients was studied for the research. The Ensemble approach produced a precision of 76.4% for the suggested model.

Nicolò, C., Périer, C., et al., (2020) suggested early-stage prediction of metastatic recurrence in breast cancer patients. They implemented the model, Random survival forest, Logistic regression, Gradient boosting, SVM, RF, and Cox techniques k-NN using a dataset of 642 patients with 21 clinicopathologic variables. The outcomes demonstrated that the mechanistic model best matched the data. The model should be evaluated using external datasets for future research.

Mosayebi, A., Mojaradi, B., Bonyadi Naeini, A. and Khodadad Hosseini, S.H., (2020) evaluated the machine learning classifier techniques for determining the relapse of breast cancer. The authors used Random Forest, Bayesian Neural Network, Artificial Neural Network, C5.0, Multilayer Perceptron, KPCA-SVM, and LVQ neural network. With 81% accuracy, the C5.0 algorithm was the best method for predicting breast cancer recurrence in the first five years.

Kabiraj, S., Akter, L., et al., (2020) analyzed the factors leading to breast cancer recurrence. Age, sex, procured inherited deformations, and skin type was found to be the main features in the prediction of breast cancer to re-occur. The Naive Bayes classifier was considered as the basis for the Bagging method, which the authors employed to estimate whether the patient might suffer a relapse of cancer or not. 73.8182% accuracy was recorded in the research process. The sample size was 275 records.

Goyal, K., Aggarwal, P. and Kumar, M., (2020) studied the diagnosis and prognosis of breast cancer at an early stage as it can be life-threatening to patients who had breast cancer earlier. They guided towards appropriate medical treatment and handling of the data in an efficient manner for the patients who suffer from breast cancer issues again. The different techniques, Generalized Regression Neural Network (GRNN), SVM, DT, NB, Feed Forward Back Propagation Neural Network (FFBPN) have been implemented by the authors to estimate the chances of the relapse of cancer. FFBPN performed the best as compared to other methods, with an accuracy rate of 85.18%.

Zain, Z.M., Alshenaifi, M., et al., (2020) used KNN, REPTree, and NB models to forecast breast cancer return. These models were built, including and excluding the feature extraction algorithm of principal component analysis (PCA). KNN without using PCA gave better accuracy than others. Further, other feature extraction algorithms other than PCA can also be used in the future to increase the performance and efficiency of the proposed models.

Chaudhuri, A.K., Sinha, D., Bhattacharya, K. and Das, A., (2020) examined that on an international scale, Breast cancer claims the lives of more women than every illness, and cancer, in general, is on the rising general has devastating impact on humanity. This disease's recurrence is one of its distinguishing features. One must proceed cautiously in the future in the search for significant breast cancer risk factors. There have been medical applications of data mining. For detecting difficulties, using a hybrid paradigm is better than relying only on a single approach. This study suggests an integrated method to eliminate Errors of type 1 and type 2 in forecasting

recurrence. Consider contradicting and essential aspects before deciding whether to use or stop them from the dataset. The possibilities for recognizing methodology are widened when critical techniques are evaluated using original and altered datasets. Results show that accuracy rises when variables are confined to those who were identified as reasonably important, and the database is updated following the removal of contradicting variables from it.

Akinnuwesi, B.A., Macaulay, B.O. and Aribisala, B.S., (2020) stated that even though the exact etiology of breast cancer (BC) is still a mystery to researchers, it is among the most prevalent malignancies in women across the world. Medical imaging and computational methods must be used in conjunction with clinical competence to accomplish this. Early analytic models based on support vector machine (SVM) and Principal Component Analysis (PCA) were developed to swiftly detect the existence of BCa in research studies. In the first phase of preprocessing, PCA was used to extract structures; nevertheless, the set of attributes was reduced even lower in the second point of the process, which was performed after the first stage. SVM was used to predict breast cancer risk and diagnoses using pre-processed multi-predicted data.

Thani, I. and Kasbe, T., (2020) intended that Breast cancer is a complex and multi-faceted illness increasing the death rates among women worldwide. It is an illness that women in previous generations were uncomfortable discussing, but women now feel comfortable discussing and sharing their opinions about this sickness. In the early phases of the sickness, there are no precise indicators that must be utilized to identify the condition, and the number of people who die from this condition is growing in mainly rural regions. The authors discussed the causes, the methods of detecting, and the many categories of breast cancer.

Roberto Cesar, M.O., German, et al., (2020) developed a model employing J48 and random forest, Naive Bayes and Naive Bayes Simple, SMO Poli-kernel, and SMORBF-Kernel classifiers to predict whether or not patients will experience a recurrence of breast cancer after undergoing surgery for the condition. WEKA data mining tool has been used to obtain the best classification algorithm and then integrated with the K-Means cluster Algorithm.

Lou, S.J., Hou, M.F., et al., (2020) examined various deep learning techniques to forecast the return of breast cancer following a breast cancer surgery after ten years. The study examined the Naive Bayesian Classifier (NBC), Cox Proportional-Hazards Regression (COX), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and K-Nearest Neighbor (KNN) to analyze

disease recurrence and to discover the determinants that can contribute to this lethal disease. The attributes considered included the preoperative quality of life, clinical data, demographic characteristics, and care quality. ANN model gave the highest accuracy rate. The distinct feature of their study was 171 additional datasets representing a large general population.

Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P. and Dhillon, S.K., (2019) developed models for identifying prognostic variables affecting relative survival rate using a range of machine learning approaches. The suggested model was built using support vector machines, decision trees, random forests, neural networks, extreme boost, and logistic regression. With 8066 entries, the study used a sizable dataset from the University Malaya Medical Centre in Kuala Lumpur, Malaysia. This study identified several relevant factors, including the categorization of cancer stage, tumor size, lymph nodes that tested positive, the total number of axillary lymph nodes removed, main therapy types, and diagnostic methods. The accuracy of the RF algorithm was higher (82.7%) than that of the other study techniques.

Chaudhuri, A.K., Sinha, D. and Thyagaraj, K.S., (2019) aimed to create a method to more accurately predict and identify a patient's chance of acquiring breast cancer again. Data for the study was sourced from the UCI Repository. It has 286 records, with 10 distinct dependent and independent properties. Decision tree and discriminant analysis were the techniques used. The findings indicated that two important factors, deg-malign (i.e., degree of malignancy) and node-caps (i.e., whether or not the tumor penetrated the lymph node capsule), influence the likelihood of breast cancer recurrence.

Tseng, Y.J., Huang, C.E., et al., (2019) discussed that based on serum tests and tumor cell surface indicators, breast cancer outcomes may be predicted. The prognosis for breast cancer may, however, be better understood by examining a variety of clinicopathological traits. The serum human epidermal growth factor receptor 2 (sHER2) was evaluated by the authors as a component of a set of clinicopathological features used to predict breast cancer metastasis. These features included logistic regression, support vector machine, random forest, and Bayesian classification algorithms. 302 people who received a diagnosis of breast cancer, had treatment for it, and underwent at least one sHER2 test between 2003 and 2016 made up the sample cohort. The classifier, based on the random forest technique, was discovered to be the most accurate at predicting breast cancer, with 75% of cases correctly predicted.

Chakradeo, K., Vyawahare, S. and Pawar, P., (2019) diagnosed the relapse of cancerous cells; different techniques were evaluated using several measures, like accuracy, efficiency, and precision. Support Vector Machine (made using RBF Kernel and Leave-One-Out (K-fold CrossValidation)), Multiple Linear Regression, and Decision Tree (built with metrics like Gini Index, Entropy, and Information Gain) are the models developed. The Support Vector Machine and K-fold Cross-Validation techniques were optimal solutions for predicting cancer relapse and the absence of recurrence. The best results were produced by SVM with RBF, which outperformed all other models.

Zeng, Z., Yao, L., et al. (2019) developed a classifier to identify breast cancer relapse using clinical and structured data from the EHR. In addition to obtaining formal clinical data from the EHR, scientists employed MetaMap for the extraction of features from clinical narratives. With these variables, authors created a model using SVM that can identify whether the patients suffering from breast cancer will have changes of recurrence or not. For the purpose of efficiently and correctly diagnosing distant breast cancer recurrences, the algorithm included aspects from standardized clinical data and unstructured clinical narratives. 1396 double-annotated subjects were utilized by the authors to train the model, while 599 double-annotated subjects were used to validate it.

Lafourcade, A., His, M., et al., (2018) discussed a prospective cohort research called E3N (Etude Epidémiologique auprès de Femmes de la Mutuelle Générale de l'Education Nationale) was started in 1990 to look into the causes of the most prevalent cancers. The three forms of recurrent events—metastasis, second primary breast cancer, and locoregional recurrence—as well as overall survival, were taken into account. A joint frailty model was used to evaluate recurring occurrences and death. The 4926 E3N cohort women initially had primary breast invasive cancer were included in the study. 1334 instances recurred during the follow-up (7. 2 years on average), and 469 females lost their lives. Smoking raised the likelihood of relapse. The odds of passing away between five and ten years following diagnosis were six and twenty, respectively.

Tapak, L., Shirmohammadi-Khorram, N., et al., (2018) used machine learning techniques to research the prognosis of survival and metastasis in cancerous patients. They focused on the fact that if early diagnosis is done and proper treatment is given to breast cancer patients, then their survival rate can be increased by more than 86%. Support vector machines (SVM), Random Forest

(RF), least-square SVM (LSSVM), AdaBoost, Naive Bayes (NB), Logistic Regression (LR), Least-square SVM (LSSVM), Adabag, and Linear Discriminant Analysis Methods for predicting the survival rate and metastasis were employed. The number of patients to be alive was around 83.4%, and 85% of patients had no metastases out of a total of 550 records. For metastasis prediction, LDA and LR had the highest accuracy of 86%.

Almuhaidib, D.A., Albusayyis, F.M., et al., (2018) focused on using the ensemble learning technique to enhance breast cancer recurrence prediction. The authors aim to build a website where physicians can enter basic factors leading to breast cancer and predict the whether the patients have a chance of recurrence of breast cancer or not. The study anticipated that the accuracy of the outcomes for the research is 0.6522 for the random forest, 0.66 for decision trees, and 0.5913 for naive Bayes. The dataset included patients from Saudi; this study can be expanded to patients from different countries as different people will have other causes and effects of breast cancer recurrence.

Zhou, H., Dong, D., et al., (2018) analyzed that computed tomography (CT) radio mic imaging is used to evaluate the distant metastatic potential in lung cancer patients. The clinical data and 485 radio mic characteristics from the pre-treatment CT images were used to build a set of features for further analysis. The most beneficial characteristics were chosen using a feature selection process (feature selection through concave minimization). Classification accuracy was 71.02 percent, and the area under the curve was 72.84 percent when utilizing just CT radio mic information in the Stochastic gradient descent (SGD) technique. The radiometric characteristics of the CT imaging obtained before treatment may be utilized to forecast distant metastases in certain cases.

Sakri, S.B., Rashid, N.B.A. and Zain, Z.M., (2018) have shown that having access to cutting-edge methods and detailed datasets can assist in predicting breast cancer recurrence at an early stage, significantly extending patients' lives. Particle swarm optimization is integrated with three well-known classifiers—K-nearest neighbour, Naive Bayes, and fast decision tree learner—for feature selection. Both clustering approaches and ensemble algorithms will be put to the test.

2.1 Summary of the reviewed literature

Table 1 shows the different data mining techniques used by the various authors and their findings.

Table 1. Summary of the existing work

S.NO.	AUTHORS	PAPER TITLE	TECHNIQUES USED	OUTCOME
1	**Ahmad, S., Ullah, T., et al., 2022**	A Novel Hybrid Deep Learning Model for Metastatic Cancer Detection.	Convolutional neural network GRU (CNN-GRU), CNN long short-term memory (CNN-LSTM), and the proposed AlexNet-GRU.	The proposed AlexNet-GRU model predicted 3% greater accuracy rate and efficiency traits when compared to existing models. More layers could have been added, and other mining techniques could have been explored.
2	**Merouane, E. and Said, A., 2022**	Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers.	Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN) and Logistic Regression techniques	Metastasis and recurrence in breast cancer patients were estimated, allowing early treatment to be given to the patients and hence saving them from metastatic relapse.
3	**Alwohaibi, M., Alzaqebah, M., et al., 2021**	A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction.	Brain Storming Optimization algorithm (BSO)	There were three main stages implemented in their research, and the best classification results with the highest accuracy were obtained in the third stage. To get more precision, the model needs to be tested on its subsets using different feature selection methods.
4	**Magboo, V.P.C. and Magboo, M.S.A., 2021**	Machine Learning Classifiers on Breast Cancer Recurrences.	Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and SVM	Recurrence of breast cancer was diagnosed with these four techniques in the patient post their surgery. It had a limitation of a small dimensional dataset being used for the research.
5	**Okagbue, H.I., Adamu, P.I., et al., 2021**	Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer.	k-nearest neighbour (kNN), Support Vector Machine (SVM), Logistic Regression, Naive Bayes (NB), Adaptive boosting (AdaBoost), Random Forest (RF), Stochastic Gradient Descent (SGD), Gradient boosting (GraBoost), Classification Tree (Tree) (LR).	Age, time duration, method of diagnosis, sex, and disease site can all be used to determine whether a breast cancer patient will live or die. Larger data sets are to be used in the future. No clarity on the range of age for Patients suffering from a relapse of cancer.
6	**Yang, P.T., Wu, W.S., et al., 2021**	Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning.	AdaBoost	To predict the recurrence of breast cancer, the authors devised a model based on clinical and numerical data.

S.NO.	AUTHORS	PAPER TITLE	TECHNIQUES USED	OUTCOME
7	**Kaur, I., Doja, M.N., et al., 2021**	An Integrated Approach for Cancer Survival Prediction Using Data Mining Techniques.	Sequence mining and Ensemble Technique	The patients were categorized by the authors as being alive or dead. The Ensemble approach yielded an accuracy of 76.4% for the suggested model. Data was collected manually, and no genetic data were used.
8	**Nicolò, C., Périer, C., et al., 2020**	Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer.	Logistic regression, Mechanistic model, Support vector machine, Cox Random survival forest, k-Nearest neighbor, Random forest, Gradient boosting	It was shown that the mechanistic classifier gave the best accuracy in predicting metastatic relapse at an early stage. More data sets could have been used.
9	**Mosayebi, A., Mojaradi, B., Bonyadi Naeini, A. and Khodadad Hosseini, S.H., 2020**	Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer.	Artificial neural network, KPCA-SVM, Random Forest, LVQ neural network, Bayesian neural network, and C5.0	The C5.0 algorithm emerged as the most effective technique for predicting the relapse of cancerous cells in the first five years, with an efficiency of 81%. Patients with less than 5 years of have been used in the study.
10	**Kabiraj, S., Akter, L., et al., 2020**	Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm.	Bagging machine learning algorithm	Age, sex, inherited genetic deformations, and skin type were significant predictors of breast cancer recurrence. Sample size should be increased with different data mining techniques using ensemble learning.
11	**Goyal, K., Aggarwal, P., and Kumar, M., 2020**	Prediction of Breast Cancer Recurrence: A Machine Learning Approach.	GRNN (Regression Neural Network), SVM, FFBPN (Feed Forward Back Propagation Neural Network), DT, and NB	Predicted the recurrence of breast cancer disease using machine learning algorithms, and FFBPN gave the best results. Combination of classifiers to be used in the future.
12	**Zain, Z.M., Alshenaifi, M., et al., 2020**	Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis	Rapid Decision Tree (REPTree), Nave Bayes, and K-Nearest Neighbour(KNN)	Machine learning algorithms use feature extraction algorithms of principal component analysis (PCA) to give better results. Better feature extraction methods are to be used.
13	**Chaudhuri, A.K., Sinha, D., Bhattacharya, K. and Das, A., 2020**	An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research.	Data Mining hybrid model	The findings suggest that accuracy increases when variables are limited to those that have been classified as relatively relevant, and the dataset is updated after contradictory variables are removed.
14	**Akinnuwesi, B.A., Macaulay, B.O. and Aribisala, B.S., 2020**	Breast cancer risk assessment and early diagnosis using Principal Component	Support Vector Machine Hybrid Model (SVM) followed by PCA	Preprocessing breast cancer data, BC-RAED can categorize patients according to risk characteristics and classify instances as either malignant

S.NO.	AUTHORS	PAPER TITLE	TECHNIQUES USED	OUTCOME
		Analysis and support vector machine techniques.		or benign based on established technical indications
15	**Thani, I. and Kasbe, T., 2020**	Breast Cancer: State–of-the-art causes and diagnosis.	Fuzzy logic, Random Forest, Evolutionary algorithms.	The authors analyzed that cancer is the most important cause of the mortality rate all over the world. Initially, it is very difficult to track the symptoms of breast cancer in women, and that is why the death rate is increasing at an alarming rate, especially in rural areas. The authors discuss the general causes, symptoms, and categories of breast cancer.
16	**Roberto Cesar, M.O., German, et al., 2020**	Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis.	J48,Random Forest, Naïve Bayes	In order to determine if breast cancer had returned in the individuals who had undergone surgery, the study utilized conventional classifiers. In order to get the desired outcomes, the WEKA tool was employed, followed by the K means clustering algorithm. In the future, a hybrid model can be built for patients who have undergone surgery.
17	**Lou, S.J., Hou, M.F.,et al., 2020**	Machine Learning Algorithms to Predict Recurrence within 10 Years after Breast Cancer Surgery: A Prospective Cohort Study.	ANN, KNN Model, SVM, Naïve Bayesian Classifier Model, Cox Proportional-Hazards Regression Model	The goal of the study was to anticipate the disease's recurrence and to find the predictors that may cause this fatal condition. The attributes considered included demographics, clinical information, treatment standards, and preoperative quality of life. Results from the ANN model had the highest accuracy. A model with two levels or multi-levels could provide better results.
18	**Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P. and Dhillon, S.K., 2019**	Predicting factors for survival of breast cancer patients using machine learning techniques.	Decision Tree, Support Vector Machine, Extreme Boost, Logistic Regression, Random Forest, and Neural Networks.	Significant variables in this study included the classification of cancer into stages, the tumor size, lymph node removal, positive tested lymph nodes, the elementary therapy types, and the methods of diagnosis. The RF algorithm offered greater accuracy than other algorithms. The study should be extended to Asian women from rural areas and can be translated into a decision-making software model.

S.NO.	AUTHORS	PAPER TITLE	TECHNIQUES USED	OUTCOME
19	**Chaudhuri, A.K., Sinha, D. and Thyagaraj, K.S., 2019**	Identification of the Recurrence of Breast Cancer by Discriminant Analysis.	Decision Tree and Discriminant Analysis	Predicted the chances for breast cancer to occur again and identified the degree of malignancy and whether the tumor has penetrated lymph nodes or not as key parameters for breast cancer recurrence. Many risk factors, like overweight, high BMI, smoking, drinking, pregnancy age, etc., have been ignored and need to be monitored.
20	**Chakradeo, K., Vyawahare, S. and Pawar, P., 2019**	Breast cancer recurrence prediction using machine learning.	Decision Tree, Multiple Linear Regression, Support Vector Machine	The best predictions for the chance of breast cancerous cells relapse came from SVM and K-fold Cross-Validation. The Ensemble technique could have given better results by constructing a software system.
21	**Tseng, Y.J., Huang, C.E., et al., 2019**	Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies.	Bayesian classification algorithms, Support Vector Machine, Logistic Regression, Random Forest	The scientists identified the serum human epidermal growth factor receptor 2 (sHER2) as a part of a group of clinicopathological characteristics that were to be utilized in conjunction to predict breast cancer metastasis in the patients. The limitation was related to the small population, and data was collected from a single institution.
22	**Zeng, Z., Yao, L., et al. 2019**	Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning.	SVM	The authors applied SVM to forecast the relapse of cancer and used MetaMap to extract clinical characteristics from HER. More data mining algorithms and hybrid models could have been used.
23	**Lafourcade, A., His, M., et al., 2018**	Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort.	E3N Cohort Model	According to the researchers, instances with high-grade, sizable tumors, axillary nodal involvement, and negative estrogen and progesterone receptors were more likely to result in mortality and recurrence. Factors like the treatment of patients along the journey and changes in the habits of the patients were ignored.
24	**Tapak, L., Shirmohammadi-Khorram, N., et al., 2018**	Prediction of survival and metastasis in breast cancer patients using machine learning classifiers.	Support Vector Machines (SVM), Least Square SVM (LSSVM), AdaBoost, Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Adabag, and Linear Discriminant Analysis (LD)	The authors investigated the metastasis and prognosis of breast cancer patients. Out of 550 data points, 83.4% of the patients were still alive, and 85% of them had no metastases.

S.NO.	AUTHORS	PAPER TITLE	TECHNIQUES USED	OUTCOME
25	**Almuhaidib, D.A., Albusayyis, F.M., et al., 2018**	Ensemble learning method for the prediction of breast cancer recurrence.	Decision trees, Random Forest, and Naive Bayes	This work aims to improve breast cancer recurrence prediction using ensemble learning techniques and to create a website that enables doctors to enter features related to a breast cancer patient and determine the chance of breast cancer recurrence. The online model system can be built further by using different combinations of mining algorithms.
26	**Zhou, H., Dong, D., et al., 2018**	Diagnosis of distant metastasis of lung cancer: based on clinical and radiomic features.	SVM with SGD	The proposed SVM with Stochastic Gradient Descent (SGD) approach had a classification accuracy of 71.02 percent, and the area under the curve was 72.84% using just CT radionics features.
27	**Sakri, S.B., Rashid, N.B.A. and Zain, Z.M., 2018**	Particle swarm optimization feature selection for breast cancer recurrence prediction.	Naive Bayes, K-Nearest Neighbor, and Decision Tree	Three well-known classifiers are combined with particle swarm optimization as feature selection to increase the prediction model's accuracy level. Feature selection and ensemble techniques to be used in the future.

3. Research Gaps

The gaps identified in the existing work are as follows:-

1) The association between the relapse of cancer in patients and their age is not clear in the previous work done. The range of age is ignored so as to specify which age group of people is suffering from a recurrence of cancer.

2) More feature extraction algorithms and different combinations of data mining techniques using the concept of ensemble learning to increase accuracy need to be explored.

3) Prediction of the stage of the disease is not focused. The stage and severity of the cancerous cells affecting the body are ignored.

4) It has been found that parameters like family history, patients' history, unhealthy lifestyle, etc., are ignored, which may also affect the accuracy of results.

5) Less work is found to be done in the direction of the effect of cancerous cells on the other parts of the body, i.e. metastasis.

4. Description of Broad Area

4.1 Data Mining Process

Data mining is the core of the knowledge discovery process. Data Mining is used to extract hidden information from the given data set. The following are the steps that make up the process of information extraction, as shown in Figure 1:-

- Data Cleaning – Noise and incorrect data are eliminated during this preprocessing stage.
- Data Integration − A comprehensive data set is created by combining data from several sources.
- Data Selection – To enable the use of data mining methods, the database is searched for the information pertinent to the analytical work.
- Data Transformation – The data is converted or transformed into a uniform format so that there is no inconsistency. Data reduction is also carried out in this step.
- Data Mining – Data mining intelligent algorithms are applied in order to extract data patterns and find hidden relationships.
- Pattern Evaluation – The exciting patterns and frequent item sets are identified.
- Knowledge Presentation – The knowledge discovered is represented through different analysis tools.

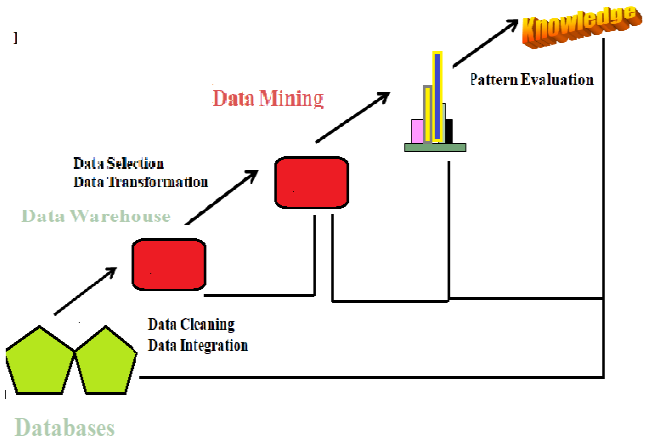

Figure 1. Data Mining Process (source: Poonia, 2003)

Figure 1 demonstrates the different data mining processes, starting with data cleaning and moving on to data integration, data selection, and data transformation before using a data mining algorithm to evaluate patterns and uncover information.

4.2 Data mining in healthcare

In the healthcare sector, patient data has been increasing exponentially today, which has led to the rise of Electronic Health Records (EHR). With data mining, healthcare organizations have optimized treatment efficiency, increased patient life, diagnose the disease at an early stage, improved customer relationship management, etc. Data mining enables physicians to treat patients more effectively and better comprehend medical tools (Ogundele, I.O., Popoola, O.L., Oyesola, O.O. and Orija, K.T., 2018). Many challenges are faced while dealing with such a large amount of data (Yahyaoui, A., Jamil, A., Rasheed, J. and Yesiltepe, M., 2019). Data mining and predictive analysis have opened new access for diagnosing and treating diseases at an early stage in the Healthcare industry. Medical data mining helps evaluate and predict early symptoms to treat any disease or condition. This reduces the money spent and saves the patients' lives because, usually, in the medical field, the decisions are made based on the doctor's experience rather than on the hidden information in the databases.

Effective data mining algorithms in the healthcare industry can further reduce hospital admissions and claim to aid better healthcare facilities (Mining, D., 2019). In cases of irregular or aberrant patterns, fraudulent medical coverage, or insurance, data mining concepts can be employed to spot fraud and abuse.

4.3 Diagnosis of Cancer

Health is the most critical asset in everyone's life. Diseases strike the human body without giving warning. Sometimes the cause of the disease is due to the family history or genetics, which no one has control over. There are different health problems like Heart Problems, Diabetes, Depression, Cancer, Blood Pressure, etc., which are very common today and can even lead to death in many cases. Cancer has become one of the most deadly diseases in the last years, leading to more than 10 million deaths in 2020. Cancer is characterized by the body's cells growing out of control and dispersing to other body regions. According to the World Cancer Report published by Cancer Research for Cancer Prevention, the financial toll of cancer is also rising daily. Lyon: The International Agency for Research on Cancer estimated that the total yearly economic cost of cancer would reach $1.16 trillion in 2020. According to World Health Organization (WHO), cases of breast cancer are 2.26 million, cancer of the lungs; it was estimated at around 2.21 million cases,

1.20 million cases of skin cancer, and 1.09 million cases of stomach cancer. About 685,000 women worldwide have died from breast cancer too far.

The incidence of the disease grows, and it impacts women of any race as women become older. Breast Cancer has more death rate in females than any other cancer of the body. Most deaths from breast cancer are caused by metastasis and recurrence of the disease (or relapse). Metastatic relapse may develop anywhere from months to decades following the original diagnosis and therapy of breast cancer (Magboo, V.P.C., and Magboo, M.S.A., 2021). Early breast cancer treatment can improve patients' chances of surviving and perhaps save their lives (Latif, M.Z., Shaukat, K., et al., 2020). One has to visit the doctor for a clinical assessment if there is any symptom or pain in the breast. The doctor examines the breast for any lumps or other signs like changes in the texture, shape, or size of the nipple. Sometimes there is the discharge of blood from the nipples. The three primary forms of breast cancer are as follows: -

1. Benign Breast Cancer – This type of cancer is not a threat to life, and it does not spread to other body parts. These cells often remain inside their origin only and do not spread outside. Because it started inside milk ducts, this cancer is also known as ductal carcinoma in situ (DCIS).
2. Malignant Breast Cancer is hazardous and can lead to death if not treated correctly. These cancer cells can spread abnormally to other body parts as well. 80% of the total patients suffering from this type of ductal cancer, and it can usually be removed with surgery.
3. Other types of Breast Cancer – This is usually rare and comprises medullary breast cancer, papillary breast cancer, tubular breast cancer, inflammatory breast cancer, and milk-producing lobules.

Recurrence was observed in roughly 30% to 40% of breast cancer patients, and approximately 10% to 15% of those patients died because of cancer metastasis or recurrence (Tseng, Y.J., Huang, C.E., et al., 2019). Global Cancer Observatory (GLOBOCAN) found that Globally, cancer is the leading cause of mortality rate, with India accounting for the majority of deaths in the developed world. Cancer patients' prognosis has been the subject of much study and experimentation by doctors and scientists (Kate, R.J. and Nadig, R., 2017) (Walczak, S. and Velanovich, V., 2018). Age, being female, genetics, a woman's reproductive history (including menarche and menopause), and hormonal state (oral contraceptives, hormone replacement therapy, radiation) are

only a few numerous risk factors for breast cancer. Menarche occurs before 11 years of age, menopause occurs after the age of 54, and abnormal radiation exposure after the age of 10 all place women in the age group for developing breast cancer (Jevtic, M., Velicki, R., et al., 2010).

Breast cancer may the spreading of cancer to other sections of the body is referred to as including weight loss without effort, seizures, shortness of breath or wheezing, swollen skin, persistent chest discomfort, exhaustion, and so on. A better cancer prognosis model may be developed with the use of data-driven prediction tools. It has been shown time and time again that data mining methods are very metastasis through the circulatory system. It may cause malignancies in the lungs, livers, bones, brains, and elsewhere. A clogged blood artery, fractured bones, and strain on the spinal cord are all possible side effects of a stroke. Women with early breast cancer are more likely to have a local recurrence within the first five years after treatment. A recurrence of breast cancer may lead to a variety of symptoms, effective in a wide range of healthcare research fields, particularly cancer treatment (Walczak, S. and Velanovich, V., 2018).

The research will aim to make sure that females with breast cancer are correctly evaluated. Women's breast cancer risk factors are multiplying at an alarming rate. Segmentation into benign and malignant subtypes is an important diagnostic strategy for breast cancer, regardless of the debate over risk assessment. This study has mainly concentrated on issues related to mammogram analysis and the development of algorithms and tools to overcome these issues to develop a classification tool with improved diagnostic accuracy to discriminate between cancer, benign and healthy breast parenchyma. The liver, muscles, lungs, brain, and bones can all be impacted if cancerous cells spread to other body parts, which can be life-threatening to the patients. It can be incredibly itchy and even dry and cracked at some point. Additional lymph nodes may be affected because of the lymphatic system. The lymphatic system, which transports healthy fluid throughout the body, may potentially disseminate cancerous tumors. Breast cancer may be detected by imaging, the knowledge of radiologists, and computer methods. The quality of the dataset's chosen features must be high sufficient to achieve an appropriate classification result.

4.4 Data mining techniques for breast cancer

Categorization is constructed to produce a prediction of the participants in a collection for data occurrences in data mining. This approach uses complicated data analysis to discover data relationships in large datasets. Medical datasets provide difficulties for pattern extortion because

of their complex properties (Ramana, B.V., Babu, M.S.P. and Venkateswarlu, N.B., 2011). Breast cancer detection and prognosis are challenging issues for researchers. Machine learning, as well as data mining, are being used more often. Approaches have followed substantial improvements in the diagnosis and prediction of breast cancer and the overall procedure of breast cancer recognition and prediction. To help with early-stage breast cancer detection, data mining is being used as an alternative means of decision support (Gupta, S., Kumar, D. and Sharma, A., 2011).

For data mining, statistical approaches and machine learning techniques are both employed. Statistics-based algorithms and machine learning algorithms are the two primary techniques for data mining. Procedures are separated into two classifications in data mining: predictive and descriptive data mining as shown in Figure 2. In a database, descriptive mining jobs are used to offer information on the general data attributes (Kharya, S., 2012).

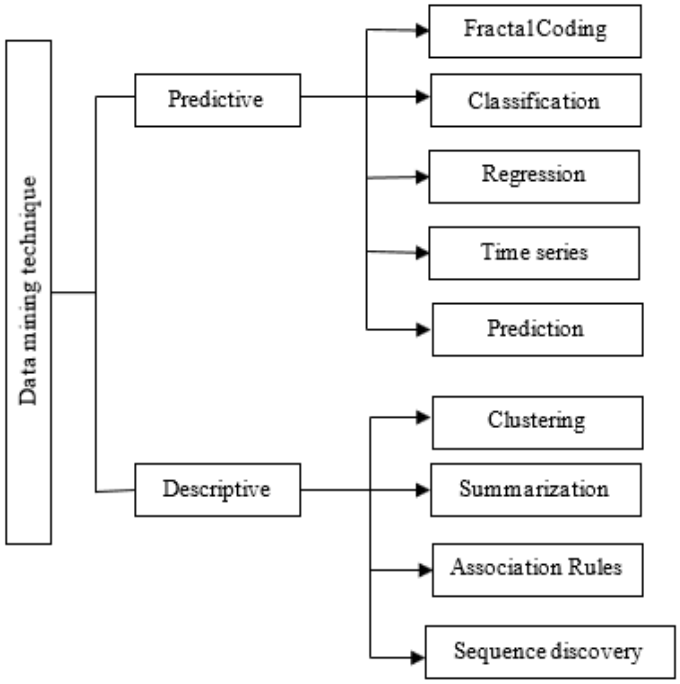

Figure 2. Data mining technique (source: Lashari, S.A., Ibrahim, R. and Senan, N., 2014).

As shown in Figure 2, the data mining techniques are classified into two broad categories of predictive and descriptive tasks. Using business intelligence or other data to analyze or predict trends is known as predictive data mining. Data categorization, association, and feature extraction are the foundations of descriptive data mining, which uses the data to describe previous behavior.

Inferences are drawn from data to make predictions, with forecasts being produced on specific values based on shapes recognized by previously observed outcomes with predictive mining tasks. Descriptive data mining gives features and descriptions for the data collection without any predefined target (Kharya, S., 2012). It is user-friendly and predictive because it learns from prior data and predicts future patterns based on past conditions. Further, data from various resources are handled, only the necessary information is retrieved, and the models may be modified by relearning past data and trends. Data mining techniques use supervised, unsupervised, and semi-supervised methods of learning. Categories are defined by the values of specific goal attributes or class attributes, which assign each instance to an appropriate category (Lashari, S.A., Ibrahim, R. and Senan, N., 2014).

5. Problem Statement

Breast cancer is the most prevalent illness that may be found in women these days, but it has lately been seen in a small number of men as well. The type of breast cancer has typically been predicted using conventional data mining approaches. Primarily, early identification of breast cancer is critical for optimal treatment since it is a life-threatening illness. Today, many patients are dying because of recurrence or relapse of cancer disease. The infection also has a significant impact on other bodily components, so an early diagnosis is necessary.

The work done in the estimation of breast cancer relapse reflects limited and low-accuracy models. There is a need to build a hybrid model with extended functionality that has more accuracy rate than the traditional mining techniques. Moreover, the inclusion of other parameters like family history, patients' history, unhealthy lifestyle, age, etc., is ignored, which may also affect the accuracy of results. Metastasis is where cancerous cells spread to other body parts such as the liver, muscles, lungs, brain, and bones. The lymphatic system, which transports healthy fluid throughout the body, may potentially disseminate cancerous tumors and can lead to the death of patients. Hence, there is a need to propose a strategy that can identify the recurrence of breast cancer and predict metastasis.

6. Research Objectives

The Research Objectives are listed as follows:-

1) To explore and analyze existing techniques for the prediction of the recurrence of breast cancer and metastasis.

2) To identify and analyze the risk factors causing Breast Cancer and determine the most common risk factors by applying feature extraction algorithms.

3) To propose an efficient hybrid model to predict the recurrence of breast cancer and metastasis in the available cancer dataset.

4) To evaluate and validate the performance of the proposed model.

7. Methodology to be adopted

The ensemble technique is a technique that may be used to enhance a classifier's accuracy. An ensemble classifier gives more accurate results than using a single classifier model. The ensemble method can be utilized to improve the accuracy of various disease prediction systems and predict the relapse or recurrence of the disease. Integrating several classifiers aims to achieve greater performance than a single classifier (Latha, C.B.C. and Jeeva, S.C., 2019). The process for the ensemble is illustrated in Figure 3.

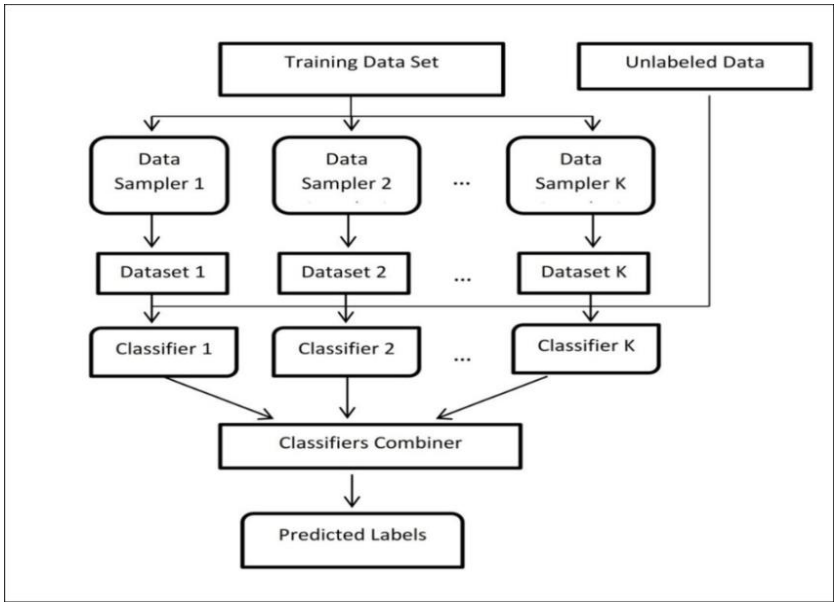

Figure 3. Ensemble process. (Source: self)

The steps which will be followed are illustrated below:-

Step 1: Dataset is considered, which contains different attributes required to predict breast cancer recurrence and metastasis.

Step 2: The dataset is preprocessed and then split. Preprocessing will include handling the missing data and removing inconsistencies if any.

Step 3. Some relevant factors of the existing dataset will be used to determine the recurrence of cancerous cells, and these factors will be trained using the feature selection algorithm. The extraction of features is often considered the most crucial phase in the categorization process. Feature extraction is the method of extracting features from an input dataset to perform computations such as detection and identification. It removes all the redundant data from the data set. In this step, the feature extraction process will be implemented.

Step 4. Various kinds of traditional classifiers, such as (Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbor) will be evaluated and executed after feature extraction in this stage.

Step 5. To specify patterns in this stage, the ensemble classifier will be utilized. This approach will combine the two classifier techniques and uses the best fitness score of each classifier to get the best overall fitness score. The ensemble approach is the most effective way since it improves the results by merging the outputs of multiple classifiers in one place.

Step 6. Recurrence of breast cancer and metastasis will be then predicted, and the accuracy of the resultant hybrid model will be compared with traditional data mining techniques.

The methodology that will be followed to build a model for the prediction of the recurrence of cancerous cells in patients who have breast cancer is illustrated in Figure 4. It will also forecast the effect of cancer on other body parts, i.e., the model will predict metastasis in the patients' diagnosed with breast cancer earlier.

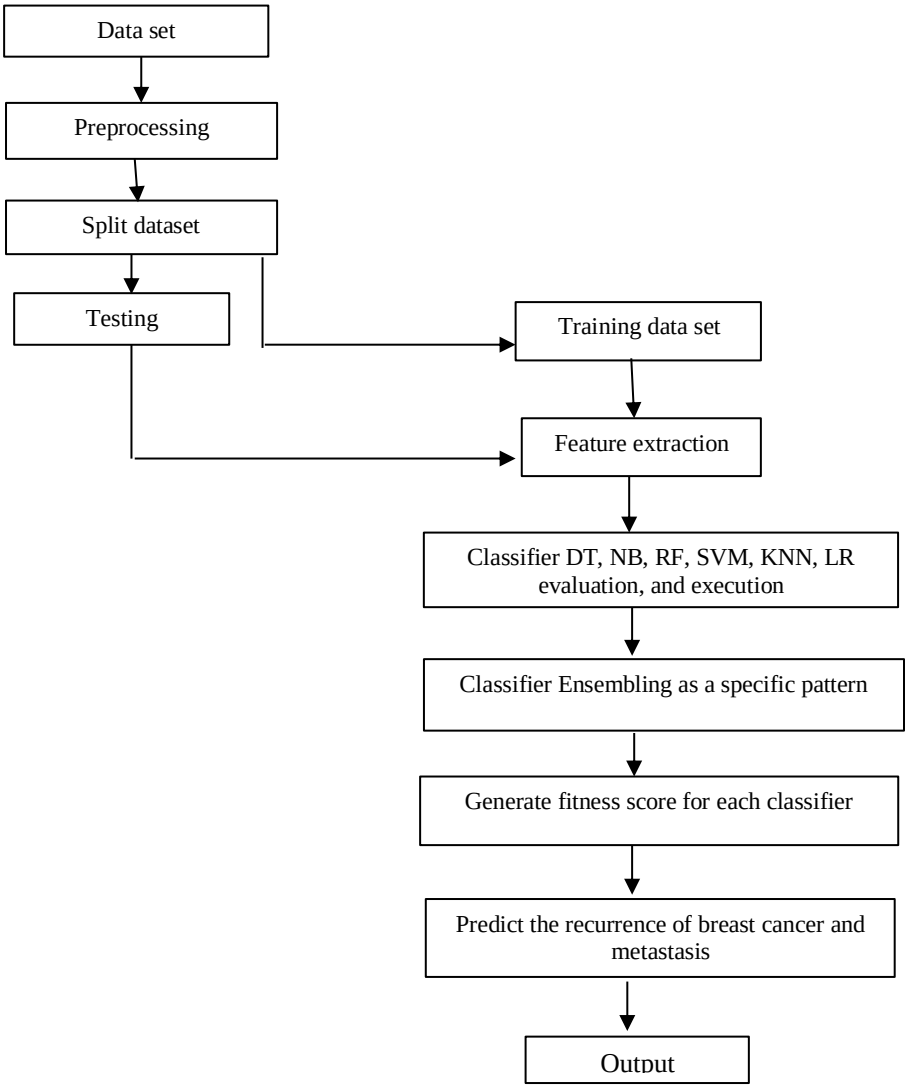

Figure 4. Proposed Methodology (Source: self)

8. Expected Outcome Of Research

The common risk factors leading to breast cancer will be identified. A model will be developed that predicts the recurrence of breast cancer and metastasis.

References

Abed, B.M., Shaker, K., Jalab, H.A., Shaker, H., Mansoor, A.M., Alwan, A.F. and Al-Gburi, I.S., 2016, November. A hybrid classification algorithm approach for breast cancer diagnosis. In *2016 IEEE Industrial Electronics and Applications Conference (IEACon)*. IEEE.

Ahmad, S., Ullah, T., Ahmad, I., Al-Sharabi, A., Ullah, K., Khan, R.A., Rasheed, S., Ullah, I., Uddin, M. and Ali, M., 2022. A Novel Hybrid Deep Learning Model for Metastatic Cancer Detection. *Computational Intelligence and Neuroscience*, *2022*.

Akinnuwesi, B.A., Macaulay, B.O. and Aribisala, B.S., 2020. Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in medicine unlocked*, *21*.

Almuhaidib, D.A., Albusayyis, F.M., Shaiba, H.A., Alzaid, M.A., Alharbi, N.G., Almadhi, R.M. and Alotaibi, S.M., 2018, April. Ensemble learning method for the prediction of breast cancer recurrence. In *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE.

Alwohaibi, M., Alzaqebah, M., Alotaibi, N.M., Alzahrani, A.M. and Zouch, M., 2022. A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction. *Journal of King Saud University-Computer and Information Sciences*, *34*(8).

Chakradeo, K., Vyawahare, S. and Pawar, P., 2019, December. Breast cancer recurrence prediction using machine learning. In *2019 IEEE Conference on Information and Communication Technology*. IEEE.

Chaudhuri, A.K., Sinha, D. and Thyagaraj, K.S., 2019. Identification of the recurrence of breast cancer by discriminant analysis. In *Emerging technologies in data mining and information security*. Springer, Singapore.

Chaudhuri, A.K., Sinha, D., Bhattacharya, K. and Das, A., 2020. An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research. *Int. J. Recent Tech. Eng*, *8*.

Doja, M.N., Kaur, I. and Ahmad, T., 2020. Age-specific survival in prostate cancer using machine learning. *Data Technologies and Applications*.

Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P. and Dhillon, S.K., 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*.

Goyal, K., Aggarwal, P. and Kumar, M., 2020. Prediction of breast cancer recurrence: a machine learning approach. In *Computational Intelligence in Data Mining*. Springer, Singapore.

Kabiraj, S., Akter, L., Raihan, M., Diba, N.J., Podder, E. and Hassan, M.M., 2020, July. Prediction of recurrence and non-recurrence events of breast cancer using bagging algorithm.

In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* IEEE.

Kaur, I., Doja, M.N., Ahmad, T., Ahmad, M., Hussain, A., Nadeem, A., El-Latif, A. and Ahmed, A., 2021. An integrated approach for cancer survival prediction using data mining techniques. *Computational Intelligence and Neuroscience*, *2021*.

Kharya, S., 2012. Using data mining techniques for diagnosis and prognosis of cancer disease. *arXiv preprint arXiv:1205.1923*.

Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M.C., Dossus, L. and Rondeau, V., 2018. Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. *BMC cancer*.

LANBARAN, N.M. and ÇELİK, E., 2021. Prediction of breast cancer through tolerance-based intuitionistic fuzzy-rough set feature selection and artificial neural network. *Gazi University Journal Of Science*.

Lashari, S.A., Ibrahim, R. and Senan, N., 2014, December. De-noising analysis of mammogram images in the wavelet domain using hard and soft thresholding. In *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*. IEEE.

Lashari, S.A., Ibrahim, R., Senan, N. and Taujuddin, N.S.A.M., 2018. Application of data mining techniques for medical data classification: a review. In *MATEC Web of conferences* (Vol. 150, p. 06003). EDP Sciences.

Latif, M.Z., Shaukat, K., Luo, S., Hameed, I.A., Iqbal, F. and Alam, T.M., 2020, June. Risk factors identification of malignant mesothelioma: A data mining based approach. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE.

Lou, S.J., Hou, M.F., Chang, H.T., Chiu, C.C., Lee, H.H., Yeh, S.C.J. and Shi, H.Y., 2020. Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study. *Cancers*.

Magboo, V.P.C. and Magboo, M.S.A., 2021. Machine Learning Classifiers on Breast Cancer Recurrences. *Procedia Computer Science*.

Merouane, E. and Said, A., 2022. Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers. *International Journal of Advanced Computer Science and Applications*, *13*(2).

Mining, D., 2019. Application of data mining techniques to predict breast cancer. *Procedia Comput. Sci*.

Mojrian, S., Pinter, G., Joloudari, J.H., Felde, I., Szabo-Gali, A., Nadai, L. and Mosavi, A., 2020, October. Hybrid machine learning model of extreme learning machine radial basis function for breast cancer detection and diagnosis; a multilayer fuzzy expert system. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE.

Mosayebi, A., Mojaradi, B., Bonyadi Naeini, A. and Khodadad Hosseini, S.H., 2020. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PloS one*.

Nicolò, C., Périer, C., Prague, M., Bellera, C., Macgrogan, G., Saut, O. and Benzekry, S., 2020. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO clinical cancer informatics*.

Ogundele, I.O., Popoola, O.L., Oyesola, O.O. and Orija, K.T., 2018. A review on data mining in healthcare. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*.

Okagbue, H.I., Adamu, P.I., Oguntunde, P.E., Obasi, E. and Odetunmibi, O.A., 2021. Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. *Health and Technology*.

Roberto Cesar, M.O., German, L.B., Paola Patricia, A.C., Eugenia, A.R., Elisa Clementina, O.M., Jose, C.O., Marlon Alberto, P.M., Fabio Enrique, M.P. and Margarita, R.V., 2020, July. Method based on data mining techniques for breast cancer recurrence analysis. In *International Conference on Swarm Intelligence*. Springer, Cham.

Sakri, S.B., Rashid, N.B.A. and Zain, Z.M., 2018. Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*.

Shukla, N., Hagenbuchner, M., Win, K.T. and Yang, J., 2018. Breast cancer data analysis for survivability studies and prediction. *Computer methods and programs in biomedicine*.

Simsek, S., Kursuncu, U., Kibis, E., AnisAbdellatif, M. and Dag, A., 2020. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert Systems with Applications*.

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O. and Poorolajal, J., 2019. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*.

Thani, I. and Kasbe, T., 2020, February. Breast Cancer: State–of-the-art causes and diagnosis. In *2nd International Conference on Data, Engineering and Applications (IDEA)*. IEEE.

Tseng, Y.J., Huang, C.E., Wen, C.N., Lai, P.Y., Wu, M.H., Sun, Y.C., Wang, H.Y. and Lu, J.J., 2019. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International journal of medical informatics*.

Walczak, S. and Velanovich, V., 2018. Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks. *Decision Support Systems*.

Wang, H., Zheng, B., Yoon, S.W. and Ko, H.S., 2018. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*.

Yahyaoui, A., Jamil, A., Rasheed, J. and Yesiltepe, M., 2019, November. A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE.

Yang, P.T., Wu, W.S., Wu, C.C., Shih, Y.N., Hsieh, C.H. and Hsu, J.L., 2021. Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. *Open Medicine*.

Zain, Z.M., Alshenaifi, M., Aljaloud, A., Albednah, T., Alghanim, R., Alqifari, A. and Alqahtani, A., 2020. Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis. *International Journal of Advances in Intelligent Informatics*.

Zeng, Z., Yao, L., Roy, A., Li, X., Espino, S., Clare, S.E., Khan, S.A. and Luo, Y., 2019. Identifying breast cancer distant recurrences from electronic health records using machine learning. *Journal of healthcare informatics research*.

Zhou, H., Dong, D., Chen, B., Fang, M., Cheng, Y., Gan, Y., Zhang, R., Zhang, L., Zang, Y., Liu, Z. and Zheng, H., 2018. Diagnosis of distant metastasis of lung cancer: based on clinical and radiomic features. *Translational oncology*.

MANAV RACHNA
Vidyanatariksha

Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research and Studies
(A deemed-to-be-university under section 3 of UGC Act 1956)

SIMILARITY VERIFICATION CERTIFICATE

Title: **Prediction of Recurrence and Metastasis in Breast Cancer**

Researcher: **Charanpreet Kaur** Registration Number: **20/PhD/054**

Supervisor: **Dr. Rosy Madaan** Department: **Faculty of Engineering and Technology**

This is to certify that the above-mentioned Short Synopsis/Thesis was scanned for similarity index through Turnitin Software on **14.12.2022** with submission ID: **1981230150** and the similarity index was found to be **9%**. The complete report running into 32 Pages, **9068** Word count with File Size 4419 **KB** is being forwarded to the Research Supervisor/DRC/DRB for review.

Dr. Rajesh Kumar 23/12/22
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

Date: 23/12/22

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Charanpreet

Research Scholar

Research Supervisor

Prediction of Recurrence and Metastasis in Breast Cancer

by Charanpreet Kaur

Submission date: 14-Dec-2022 10:36PM (UTC+0530)
Submission ID: 1981230150
File name: Prediction_of_Recurrence_and_Metastasis_in_Breast_Cancer.doc (458K)
Word count: 9068
Character count: 51231

Prediction of Recurrence and Metastasis in Breast Cancer

ORIGINALITY REPORT

Excluding matches less than 14 words

Dr. Rajesh Kumar 23/12/22
University Librarian
Dr. O.P. Jindal Central Library
Manav Rachna International Institute of Research & Studies Faridabad

9%	9%	9%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	iift.ac.in Internet Source	1%
2	academic-accelerator.com Internet Source	1%
3	www.hindawi.com Internet Source	1%
4	shodh.inflibnet.ac.in:8080 Internet Source	1%
5	www.neuroquantology.com Internet Source	1%
6	link.springer.com Internet Source	1%
7	doctorpenguin.com Internet Source	1%
8	www.mecs-press.org Internet Source	<1%
9	pubmed.ncbi.nlm.nih.gov Internet Source	<1%

10	www.researchgate.net Internet Source	<1 %
11	Katrina I. Sundus, Bassam H. Hammo, Mohammad B. Al-Zoubi, Amal Al-Omari. "Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset", Informatics in Medicine Unlocked, 2022 Publication	<1 %
12	Kaustubh Chakradeo, Sanyog Vyawahare, Pranav Pawar. "Breast Cancer Recurrence Prediction using Machine Learning", 2019 IEEE Conference on Information and Communication Technology, 2019 Publication	<1 %
13	bbrc.in Internet Source	<1 %
14	search.trdizin.gov.tr Internet Source	<1 %
15	"Computational Intelligence in Data Mining", Springer Science and Business Media LLC, 2020 Publication	<1 %
16	jebas.org Internet Source	<1 %

17	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2022 Publication	<1 %
18	pdfs.semanticscholar.org Internet Source	<1 %
19	www.eurekaselect.com Internet Source	<1 %
20	"Proceedings of Third International Conference on Communication, Computing and Electronics Systems", Springer Science and Business Media LLC, 2022 Publication	<1 %
21	care.diabetesjournals.org Internet Source	<1 %

Exclude quotes On

Exclude bibliography On

Exclude matches < 14 words

synopsis_updated_11_12_22

by charan kaur

General metrics

62,474	9,321	611	37 min 17 sec	1 hr 11 min
characters	words	sentences	reading time	speaking time

Score

This text scores better than 87% of all texts checked by Grammarly

Writing Issues

337	28	309
Issues left	Critical	Advanced

Plagiarism

56 sources

7% of your text matches 56 sources on the web or in archives of academic publications